Latent Semantic Analysis as Method for Automatic Question Scoring

David Tobinski[1] and Oliver Kraft[2]

[1] Universität Duisburg Essen, Universitätsstraße 2, 45141 Essen
david.tobinski@uni-due.de,
WWW home page: www.kognitivismus.de
[2] Universität Duisburg Essen, Universitätsstraße 2, 45141 Essen

Abstract. Automatically scoring open questions in massively multiuser virtual courses is still an unsolved challenge. In most online platforms, the time consuming process of evaluating student answers is up to the instructor. Especially unexpressed semantic structures can be considered problematic for machines. Latent Semantic Analysis (LSA) is an attempt to solve this problem in the domain of information retrieval and can be seen as general attempt for representing semantic structure. This paper discusses the rating of one item taken from an exam using LSA. It is attempted to use documents in a corpus as assessment criteria and to project student answers as pseudo-documents into the semantic space. The result shows that as long as each document is sufficiently distinct from each other, it is possible to use LSA to rate open questions.

Keywords: Latent Semantic Analysis, LSA, automated scoring, open question evaluation

1 Introduction

Using software to evaluate open questions is still a challenge. Therefore, there are many types of multiple choice tests and short answer tasks. But there is no solution available in which students may train their ability to write answers to open questions, as it is required in written exams. Especially in online courses systems (like Moodle), it is up to the course instructor to validate open questions herself.

A common method to analyze text is to search for certain keywords, as it is done by simple document retrieval systems. This method can not take into account that different words may have the same or a similar meaning. In information retrieval this leads to the problem, that potentially interesting documents may not be found by a query with too few matching keywords. Latent Semantic Analysis (LSA, Landauer and Dumais 1997) faces this problem by taking the higher-order structure of a text into account. This method makes it possible to retrieve documents which are similar to a query, even if they have only a few keywords in common.

Considering this problem in information retrieval to score an open question seems to be a similar problem. Exam answers should contain important keywords, but contain their own semantic structure also. This paper attempts to rate a student's exam answer by using LSA. For that a small corpus based upon the accompanying book of the course "Pädagogische Psychologie" (Fritz et al. 2010) is manually created. It is expected that it is in general possible to rate questions this way. Further it is of interest what constraints have to be taken into account to apply LSA for question scoring.

2 Latent Semantic Analysis

LSA was described by Deerwester et al. 1990) as a statistical method for automatic document indexing and retrieval. Its advantage to other indexing techniques is that it creates a *latent* semantic space. Naive document retrieval methods search for keywords shared by a query and a corpus. They have the disadvantage that it is difficult or even impossible to find documents if the request and a potentially interesting document have a lack of shared keywords. Contrary to this, LSA finds similarities even if query and corpus have few words in common. Beside its application in the domain of Information Retrieval, LSA is used in other scientific domains and is discussed as theory of knowledge acquisition (1997).

LSA is based upon the Vector Space Model (VSM). This model treats a document and its terms as a vector in which each dimension of the vector represents an indexed word. Multiple documents are combined in a *document-term-matrix*, in which each column represents a document and rows represent a terms. Cells contain the term frequency of a document (Deerwester et al. 1990).

A matrix created this way may be weighted. There are two types of weighting functions. Local weighting is applied to a term i in document j and global weighting is the terms weighting in the corpus. $a_{ij} = local(i,j) * global(i)$, where a_{ij} addresses a cell of the document-term-matrix (Martin and Berry 2011). There are several global and local weight functions. Since Dumais attested LogEntropy to improve retrieval results better than other weight function (Dumais 1991), studies done by Pincombe (2004) or Jorge-Botana et al. (2010) achieved different results. Although there is no consensus about the best weighting, it has an important impact to retrieval results.

After considering the weighting of the document-term-matrix, Singular Value Decomposition (SVD) is applied. SVD decomposes a matrix X into the product of three matrices:

$$X = T_0 S_0 D_0^T \tag{1}$$

Component matrix T_0 contains the derived orthogonal term factors, D_0^T describes the document factors and S_0 contains singular values, so that their product recreates the original matrix X. By convention, the diagonal matrix S is arranged in descending order. This means, the lower the index of a cell, the more information is contained. By reducing S from m to k dimensions, the

product of all three matrices ($\widehat{X}$) is the best approximation of X with k dimensions. Choosing a good value for k is critical for later retrieval results. If too many dimensions remain in S, unnecessary information will stay in the semantic space. Choosing k too big will remove important information from the semantic space (Martin and Berry 2011).

Once SVD is applied and the reduction done, there are four common types of comparisons, where the first two comparisons are quite equal: (i) Comparing documents with documents is done by multiplying D with the square of S and transposition of D. The value of cell $a_{i,j}$ now contains the similarity of document i and document j in the corpus. (ii) The same method can be used to compare terms with terms. (iii) The similarity of a term and a document can be taken from the cells of $\widehat{X}$. (iv) For the purpose of information retrieval, it is important to find a document described by keywords. According to the VSM keywords are composed in a vector, which can be understood as a query (q). The following formula projects a query into semantic space. The result is called *pseudo-document* (D_q) (Deerwester et al. 1990):

$$D_q = q^T T S^{-1} \tag{2}$$

To compute similarity between documents and the pseudo-document, consine similarity is generally taken (Dumais 1991). In their studies Jorge-Botana et al. (2010) found out that Euclidean distance performs better than cosine similarity.

3 Application configuration

To verify if LSA is in general suitable for valuating open questions, students answers from psychology exam in summer semester 2010 are analyzed. The exam question requires to describe, how a text can be learned by using the three cognitive learning strategies memorization, organization and elaboration. Each correct description is rated with two points. A simple description is enough to answer the question correctly, it is not demanded to transfer knowledge by giving an example. For the evaluation brief assessment criteria are available, but due to the short length of the description of each criterion new criteria are created by using the accompanying book of the course as mentioned above.

For the assessment a corpus is created, where each document is interpreted as an assessment criterion, which is worth a certain number of points. This way quite small corpora are created. For example, if a question is worth four points the correlating corpus contains exact four documents and only a few hundred terms, sometimes even less. To reduce noise in the corpus a list of stopwords is used. Because the students answers are short in length, stemming is used in this application. Beside using stemming and a list of stopwords, the corpus is weighted. Pincombe (2004, 17) showed that for a small number of dimensions BinIDF weighting has a high correlation to human ratings. Since the number of dimensions is that low (see below) and a human rating is taken as basis for the evaluation of LSA in this application, the used corpus is weighted by BinIDF.

All calculations are done by using GNU R statistical processing language using "lsa"[3] library provided by CRAN. The library is based upon SVDLIBC[4] by Doug Rhode. It implements multiple functions to determine the value of k. The example below was created by using *dimcalc_share* function with a threshold of 0.5, which sets $k = 2$. As consequence matrix S containing singular values is reduced to two dimensions.

Most students answers in the exam are rated with the maximum points. For this test 20 rated answers are taken, as in the exam most of them achieved the full number of points. The answers are of varying length, the shortest ones contain just five to six words, while the longest consist of two or three sentences with up to thirty or more words. Each of the chosen answers contain a description for all three learning strategies, answers with missing descriptions are ignored.

The evaluation done by the lecturers is used as template to evaluate the results of LSA. It is expected, that these answers have a high similarity to its matching criterion, represented by the documents. The rated answers are interpreted as a query, by using formula (2) the query is projected into the corpus as a pseudo-document and because of their length they be near to the origin of the corpus. To calculate the similarity between the pseudo-documents and the documents, cosine similarities is used.

4 Discussion

Figure 1 (a) shows the corpus with all three assessment criteria (0_Memorization, 1_Organization, 2_Elaboration). It is noticeable that the criterion for memorization lies closer to the origin than the other two criteria. This is a result of the relatively short length of the document which is taken as criterion for memorization. If the similarity between this and the other criteria is calculated, one can see that this is problematic. Document 1_Organization and 2_Elaboration have a cosine similarity of 0.08, so they can be seen as very unequal. While 0_Memorization and 1_Organization have an average similarity of 0.57, criteria 0_Memorization and 2_Elaboration are very similar with a value of 0.87. Therefore and because of the tendency of pseudo-documents to lie close to the origin, it can be expected that using cosine similarity will not be successful. The assessment criterion for the descriptions of the memorization strategy overlaps the criterion for the elaboration strategy.

Looking at precision and recall values proofs this assumption to be correct for the corpus plotted in Figure 1 (a). The evaluation of the answers achieves a recall of 0.62, a precision of 0.51 and an accuracy of 0.68. Although the threshold for a correct rating is set to 0.9, both values can be seen as too low to be used for rating open questions. Since the two criteria for memorization and elaboration

[3] http://cran.r-project.org/web/packages/lsa/index.html

[4] http://tedlab.mit.edu/ dr/SVDLIBC/ This is a reimplementation of SVDPACKC written by Michael Berry, Theresa Do, Gavin O'Brien, Vijay Krishna and Sowmini Varadhan (University of Tennessee).

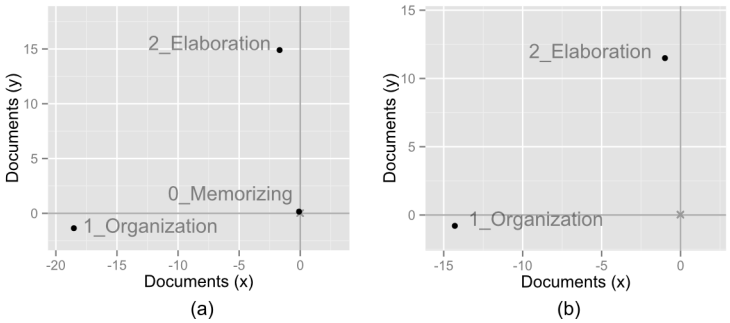

Fig. 1. Figure 1 (a) shows the corpus containing all three assessment criteria. It is illustrated that document 0_Memorizing lies close to the origin. Figure 1 (b) shows the corpus without the document 0_Memorizing. In Figure (a) and (b) the crosses close to the origin mark the positions of the 20 queries.

have a high similarity, a description for one of them gets a high similarity for both criteria. This causes the low precision values for the evaluation.

Figure 1 (b) illustrates the corpus without the document, which is used as criterion for the memorization strategy. Comparing both documents shows a similarity of 0.06. By removing the problematic document from the corpus, the similarity of the students answers to the assessment criterion for elaboration can be calculated without being overlapped by the criterion for memorization. Using this corpus for evaluation improves recall to 0.69, precision to 0.93 and accuracy to 0.83.

If one compares both results, it is remarkable that precision as a qualitative characteristic improves to a high rate, while recall stays at an average level. This means in the context of question rating that answers correctly validated by LSA are very likely rated positive by a human rater. Although LSA creates a precise selection of correct answers, recall rate shows that there are still some positive answers missing in the selection. The increase of accuracy from 0.68 to 0.83 illustrates that the number of true negatives increases by using the second corpus.

5 Conclusion and Future Work

The results of the experiment are encouraging and the general idea of using LSA to rate open questions is functional. The approach of using documents as assessment criterion and project human answers as pseudo-documents into the semantic space constructed by LSA is useful. LSA selects correct answers with a high precision, although some positive rated answers are missing in the selection. But the application shows that some points need to be considered.

All assessment criteria have to be sufficient distinct from each other and should be of a certain length, if cosine similarity is used. As the criterion for

rating the elaboration descriptions shows, it is important that no criterion is overlapped by another. Without considering this, sometimes it is impossible to distinguish which criterion is the correct one. Having a criterion overlapping another one leads to the problem that both criteria get a high similarity, which raises the number of false positives and reduces the precision of the result. This is a mayor difference between the application of LSA as an information retrieval tool or for scoring purposes.

Concerning the average recall value, it is an option to examine the impact of a synonymy dictionary in futher studies. In addition, our result shows that BinIDF weighting works well for a small number of dimensions, as Pincombe (2004) described.

For future work, we plan to use this layout in an online tutorial to perform further tests in winter semester 2013/14. The tutorial is designed as massively multiuser virtual course and will accompany a lecture in educational psychology, which is attended by several hundred students. It will contain two items to gain more empirical evidence and experience with this application and its configuration. To examine the impact on learners long-term memory will be subject to further studies.

References

Deerwester, S., Susan T. D., Furnas, G. W., Landauer, T. K., Harshman, R.: Indexing by latent semantic analysis. Jounal of the American Society For Information Science 41, 391–407 (1990)

Dumais. S. T.: Improving the retrieval of information from external sources. Behavior Research Methods 23, 229–236 (1991)

Fritz, A., Hussy, W., Tobinski, D.: Pädagogische Psychologie. Reinhardt, München (2010)

Jorge-Botana, G., Leon, J. A., Olmos R., Escudero I.: Latent Semantic Analysis Parameters for Essay Evaluation using Small-Scale Corpora. Journal of Quantitative Linguistics 17, 1–29 (2010)

Landauer, T. K., Dumais, S. T.: Solution to Plato's problem : The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, 104, 211-240 (1997)

Landauer, T. K., McNamara, D. S., Dennis, S., Kintsch, W.: Handbook of Latent Semantic Analysis. Routledge, New York and London (2011)

Martin, D. I., Berry, M. W.: Mathematical Foundations Behind Latent Semantic Analysis. Landauer et al., Handbook of Latent Semantic Analysis, 35–55 (2011)

Pincombe, B.: Comparison of human and latent semantic analysis (LSA) judgments of pairwise document similarities for a news corpus (2004)